van der Horst, E., Okuno, Y., Bender, A. and Ijzerman, A.P.	Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner	**J. Chem. Inf. Model.**	49	348-60	2009
Tsuchiya, S., Tachida, Y., Segi-Nishida, E., Okuno, Y., Tamba, S., Tsujimoto, G., Tanaka, S. and Sugimoto, Y.	Characterization of gene expression profiles for different types of mast cells pooled from mouse stomach subregions by an RNA amplification method	**BMC Genomics**		10 : 35	2009
Ruike, Y., Ichimura, A., Tsuchiya, S., Shimizu, K., Kunimoto, R., Okuno, Y., and Tsujimoto, G.	Global correlation analysis for micro-RNA and mRNA expression profiles in human cell lines	**J. Hum. Genet.**	53	515-23	2008
Kawanishi, H., Matsui, Y., Ito, M., Watanabe, J., Takahashi, T., Nishizawa, K., Nishiyama, H., Kamoto, T., Mikami, Y., Tanaka, Y., Jung, G., Akiyama, H., Nobumasa, H., Guilford, P., Reeve, A., Okuno, Y., Tsujimoto, G., Nakamura, E. and Ogawa, O.	Secreted CXCL1 is a potential mediator and marker of the tumor invasion of bladder cancer	**Clin. Cancer Res.**	14	2579-87	2008
Takano, H., Nakazawa, S., Okuno, Y., Shirata, N.,Tsuchiya, S., Kainoh, T., Takamatsu, S., Furuta, K., Taketomi, Y., Naito, Y., Takematsu, H., Kozutsumi, Y., Tsujimoto, G., Murakami, M., Kudo, I., Ichikawa, A., Nakayama, K., Sugimoto, Y. and Tanaka, S.	Establishment of the culture model system that reflects the process of terminal differentiation of connective tissue-type mast cells	**FEBS Lett.**	582	1444-50	2008

Okuno, Y.	In silico drug discovery based on the integration of bioinformatics and chemoinformatics	**YAKUGAKU ZASSHI**	**128 (11)**	1645-51	2008
藪内 弘昭, 奥野 恭史	ケミカルゲノミクス情報を用いた新規リガンド探索手法	**SAR News**	14	2-6	2008
Inoue, T., Adachi, H., Murakami, S., Takano, K., Matsumura, H., Mori, Y., Fukunishi, Y., Nakamura, H., Kinoshita, T., Nakanishi, I., Okuno, Y., Minakata, S., Shimojo, S., Sakata, T.	New progress in crystallization technology of membrane protein and introduction of pharamaceutical innovation value chain	**YAKUGAKU ZASSHI**	**128 (4)**	497-505	2008
新島　聡, 奥野　恭史	ケミカルゲノミクスに基づくインシリコ創薬	**日薬理誌**	133	173	2009
Niijima, S. and Okuno, Y.	Laplacian linear discriminant analysis approach to unsupervised feature selection	**IEEE/ACM Trans.Comput. Biol.Bioinformatics.**	IEEE computer Society Digital Library.		2007
Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimto, R. and Feng, C.	GLIDA: GPCR-ligand database for chemical genomics drug discovery-Database and tools update	**Nucleic Acids Res**	36	D907-12	2008
Kitajima, M., Minowa, Y., Matsuda, H. and Okuno, Y.	Compound-transporter interaction studies using canonical correlation analysis	**Chem-Bio Inform J.**	7	24-34	2007

Yamamoto, H., Takematsu, H., Fujinawa, R., Naito, Y., Okuno, Y., Tsujimoto, G., Suzuki, A. and Kozutsumi, Y.	Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for glycan biosynthesis enzyme gene	**PLoS ONE**	2	e1232	2007
Ikeda, A., Miyazaki, T., Kakizawa, S., Okuno, Y., Tsuchiya, S., Myomoto, A., Saito, SY., Yamamoto, T., Yamazaki, T., Iino, M., Tsujimoto, G., Watanabe, M. and Takeshima, H.	Abnormal features in mutant cerebellar Purkinje cells lacking junctophilins	**Biochem. Biophys. Res. Commun.**	363	835-9	2007
Yamazaki, T., Sasaki, N., Nishi, M., Yamazaki, D., Ikeda, A., Okuno, Y., Komazaki, S., and Takeshima, H.	Augmentation of drug-induced cell death by ER protein BRI3BP	**Biochem. Biophys. Res. Commun.**	362	971-5	2007
奥野　恭史	ケミカル・バイオ情報に基づく創薬インフォマティクス研究	**Pharma VISION NEWS**	9	13-16	2007
Naito, Y., Takematsu, H., Koyama, S., Miyake, S., Yamamoto, H., Fujinawa, R., Sugai, M., Okuno, Y., Tsujimoto, G., Yamaji, T., Hashimoto, Y., Itohara, S., Kawasaki, T., Suzuki, A., Kozutsumi, Y.	Germinal center marker GL7 probes activation-dependent repression of N-glycolylneuraminic acid, a sialic acid species involved in the negative modulation of B cell activation.	**Mol. Cell Biol.**	27	3008-22	2007
Zhu, S., Okuno, Y., Tsujimoto, G. and Mamitsuka, H.	Application of a new probabilistic model for mining implicit associated cancer genes from OMIM and Medline	**Cancer Inform**	2	361-71	2006

Osada, S., Naganawa, A., Misonou, M., Tsuchiya, S., Tamba, S., Okuno, Y., Nishikawa, J., Satoh, K., Imagawa, I., Tsujimoto, G.., Sugimoto, Y., Nishihara, T.	Altered gene expression of transcriptional regulatory factors in tumor marker-positive cells during chemically induced hepatocarcinogenesis.	**Toxicol. Lett.**	167	106-13	2006
Tsuchiya, S., Okuno, Y., Tsujimoto, G..	MicroRNA: biogenetic and functional mechanisms and involvements in cell differentiation and cancer.	**J. Pharmacol Sci.**	101(4)	267-70	2006
Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., Tsujimoto, G..	GLIDA: GPCR-Ligand database for chemical genomic drug discovery	**Nucleic Acids Research.**	34	D673-7	2006

医薬品安全性に関する文献情報自動抽出システムの考案

天野 博夫　金子 周司
京都大学大学院薬学研究科生体機能解析学分野

A newly devised text search system for adverse drug reactions.

Amano Hiro　Kaneko Shuji
Department of Molecular Pharmacology, Graduate School of Pharmaceutical Sciences, Kyoto University.

In recent years, drug safety has become a major issue for those engaged in medical care. Although medical literatures are on a watch list for drug safty matter, it is a tremendous task to sort enormous amount of text information. We devised an exhaustive text search system for specialized use in pharmacovigilance termed TSADR (Text-search System for Adverse Drug Reaction), which is made of two medical vocabularies (DN-INDI list and RN list) and a Perl script file. TSADR extracts sentences involving topics relevant to the drug safety from PubMed abstracts, creates HTML files which show extracted texts with drug names and adverse reaction names color-coded on pages in a web browser and assist searchers to discriminate important items. TSADR is now developing toward the practical use for text analysis in pharmacovigilance to identify and anticipate adverse reactions resulting from drug use.

Keywords: adverse drug reaction, literature information

1. 研究の背景と目的

製薬企業のグローバル化が進行し、新薬の開発・供給体制の迅速化が図られる一方で、医療関係者共通の重要問題である薬の副作用等、安全性に関する情報の収集・伝達体制の整備は必ずしも進んでいない。文献情報の収集・解析により副作用の発生を早期に検知あるいは予測するためには、質的なばらつきが大きい大量のテキストソースを網羅的に検索して医薬品の安全性に特化した情報を選別・抽出する作業が必要であり、これは通常のキーワード検索では事実上不可能である。本研究においては、PubMedアブストラクトを対象として、医薬品の安全性に関する情報の選別・抽出を補助するシステムの構築を目的とした。

2. 研究方法

米国FDAからAERS(Adverse Event Reporting System)データベース[1]2004年第一四半期から2005年第二四半期まで1年半分のASCIIデータファイルをダウンロードし、"DRUGNAME"、"INDI-PT"、"PT"各フィールドのデータから医薬品名とその適応を関連させたリスト(DN-INDI list)および有害反応名のリスト(RN list)を作成した。これらを辞書としてPubMedアブストラクトから医薬品名と有害反応名が同時に記載されているセンテンスを抜き出し、ヒットした用語の認識性を色別表示により改善するPerlスクリプト(TSADR)を作成した。適応名の記載が同一のセンテンス内にあれば、これも表示させた。TSADRの基本的な動作の概要をフローチャート(図1)に示す。PubMedアブストラクトはLimits設定フォームにおいてonly items with abstracts, English, Humansの3つの制限のみを設定し、キーワードを入力せずに取得した500件分のテキストを検索対象の1単位とした。TSADRにより抽出されたセンテンスを含むアブストラクトを読み、医薬品の副作用の記載が正しく抽出・表示されているかを検証した。1単位のテキストサンプルに対する一回のオペレーションの抽出率(500件中何件のアブストラクトが抽出されたか)および正解率(抽出されたアブストラクトの何%に医薬品の安全性に関する内容が記載されていたか)をシステムの評価基準とした。

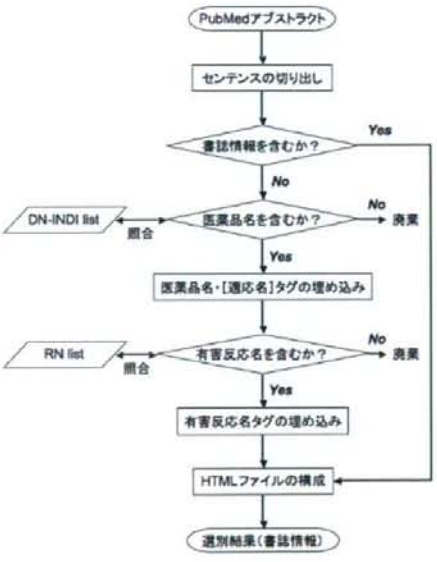


図1 TSADR基本動作のフローチャート

3. 研究成績と考察

図2は、上記研究方法に取得条件を示した、ヒトに関する英文アブストラクト付きPubMedアイテムの年間エントリー数の推移を示したものである。年を追ってエントリー数の増加が認められるが、2005年のデータを参考にすると、このコーパスから網羅性を重視して

必要なアイテムを選別していくには、一日平均800ないし1000件を処理する必要があり、その実行には豊富かつ高水準の労働力、または自動化による補助が必須であると考えられる。

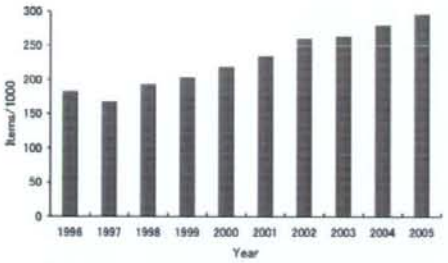


図2 PubMedアイテムの年間エントリー数
研究方法に記載の取得条件に合致する、ヒト関連の英文アブストラクト付きPubMedアイテム年間エントリー数の推移

本研究において医薬品の安全性に関連する語彙のソースとして用いたAERSデータベースは、米国FDAが収集・管理している、製薬企業からの義務報告と医療従事者・患者およびその家族からの自発報告を総合した巨大な副作用データベースであり、四半期分(収載報告件数8-9万)ごとに半年遅れの生データがASCIIまたはSGMLファイルとしてFDAのサイトから入手できる。今回は2004年および2005年上半期の一年半分のレコードをベース語彙リストのソースとして用いた。AERSデータベースはリレーショナルデータベースの構造を持つため、医薬品名と適応名を対応させて取得できる大きなメリットがある。初期システムにおいては、3813の医薬品名(drug name)、3824種類の適応(indication)をDN-INDI listに、9712種類の有害反応名(reaction name)をRN listに収載した。

初期システムを用いて、日本時間06年5月29日に取得したPubMedアブストラクト500件(11,924センテンス)より74件(138センテンス)が抽出された(抽出率14.8%)。74件中、有害反応名が正しく表示されていたものが26件(正解率35.1%)、32件では医薬品名と直接には無関係な反応がヒットし、16件は適応症が有害反応として誤って表示されていた。誤った選別のパターンとしては、"glucose"、"oxygen"、金属イオン等の生体成分が医薬品名として拾われて起こる事例や"alcohol"、"antibiotic"、"chemotherapy"等、医薬品分類名に関して誤った選別が起こる例が多く認められた。前者のパターンに関しては、隣接する単語との関連から医薬品名としての取捨を判断するフィルターをスクリプトに加えて対応し、後者のパターンはDN-INDI listの適応エントリー数を増やす手段で対応した。システムの構造上、語彙リストの医薬品名、有害反応名のレコード数を増やせば抽出率は上昇し、検索の網羅性に関しては有利に働く一方で、副作用以外の医薬品名と有害反応名の組み合わせを拾う可能性も高くなり、正解率が下がれば人間による最終的な選別操作の負担が大きくなる。これに対して、医薬品を投与する原因となる病態などの名称、すなわち適応名の語彙を増やすことは、誤った選別を抑制し、検索の精度を上昇させる。

上記対応を施したシステムを、新たに(日本時間06年6月14日)取得したシステムトレーニング用テキスト(STTXT)に適用し、その結果を基に語彙リストファイルを修正する作業を繰り返した。また、誤って選別されたアブストラクトに癌・腫瘍関係の雑誌のものが多かったことから、癌・腫瘍関係語彙用テキスト(CTBTXT: PubMedからSubsetsのLimitにCancerを設定して取得した)を用いたトレーニングも行った。これらのトレーニングによって最適化された語彙リストを用いて、両トレーニングテキスト自身を検索した最終的な成績(正解件数／抽出件数)はSTTXTが35/62(正解率56.5%)、CTBTXTが42/64(正解率65.6%)であった。必要な場合にはスクリプトの修正も行った。

本システムの実用化に向けて、システムパフォーマンスに対する上記トレーニングの有効性を検討する目的で、新たに取得したテキスト(日本時間06年8月1日)をトレーニング前のシステムTSADR-originalとトレーニング後のシステムTSADR-trainedを用いて解析し、得られた成績を比較した。TSADR-original、TSADR-trainedそれぞれの成績(正解件数／抽出件数)は14/44(正解率31.8%)および22/54(正解率40.7%)と算出され、抽出率、正解率ともにトレーニングの有効性が認められた。

一方、選別されるべきセンテンスの拾い漏れがどの程度起こっているかの予備的検討として、医薬品文献情報の有力サイトである英国のNational electronic Library for Medicines[3]に最近(Date Published 12/04/2006-24/08/2006)ピックアップされた項目のうち副作用情報に分類される120レコードを対象にTSADRによる重要文献の抽出漏れを検討した。120件中抽出されなかったレコードは14件であった(抽出率88.3%)。抽出漏れの原因としては、DN-INDI listに医薬品名が収載されていなかったものが11件、RN listに有害反応名が収載されていなかったものが5件、2件は医薬品名、有害反応名ともリストに記載はあったが、別個のセンテンス中に記載されていたためヒットしなかった。

医薬品の安全性に関する情報は、新薬に関してその重要性が特に大きく、本検索システムの実用性は医薬品名を主とする語彙リストのアップデート状況に強く依存する。今後、本システムの網羅性および選別性をさらに向上させるために、語彙リストの補強・改訂を自動化する方法を検討する予定である。

参考文献

[1] Adverse Event Reporting System(AERS).http://www.fda.gov/cder/aers/default.htm.
[2] Entrez PubMed.http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.
[3] National electronic Library for Medicines.http://www.druginfozone.nhs.uk/home/default.aspx.

分子薬理学的知識を記述する新たな三項関係データベースの開発

伊藤 悦子　金子 周司
京都大学大学院薬学研究科

The development of novel drug database describing molecular pharmacological knowledge

Ito Etsuko Kaneko Shuji
Graduate School of Pharmaceutical Sciences, Kyoto University

We have developed a novel drug database describing sites and modes of action from molecular pharmacoloical point of view. About 18,000 drug names were collected from JAPIC, JAN, ATC, KEGG and MeSH databases and rearranged to a synonym table listing 2,800 active chemical compounds. The sites of action were assigned to one or few target molecules described with MeSH descriptors and RefSeq genes. The mode of action for each interaction was described with simle descriptors, such as 'inhibition' and 'activation'. As a result, the modes of drug ation were assigned to approximately 70% of active compounds to 500 target molecules. The molecular pharmacology database will be of useful for constructing medical ontology.

Keywords: pharmacological action, ligandp, ontology, drug name

1. はじめに

生命情報学の発展により,生体分子や化合物情報の網羅的なデータベースが整いつつある。これら物質データベースが完成の域へ近づくにつれ,それぞれの関係や相互作用を記述する統合的なデータベースが知識基盤として構築されつつある[1,2]。

医薬品の作用は,究極的には活性成分である化学物質と生体分子との相互作用に単純化できる[3]。医薬品の全身作用を収録する添付文書や副作用情報は電子化が進んでいるが,薬物の分子作用点や作用メカニズムに関する薬理学的知識について,単純化した相互作用として整理したデータベースは数多くない[4,5]。特に,日本語で記述される製剤名や商品名などの多様な表記をカバーし,医療文書のテキストマイニングに耐えうる網羅的な記述子を備えた薬理作用データベースはまだない。

本研究ではまず,医薬品として用いられる化学物質に対して付与される製剤名や商品名に加え,生体分子を記述する際に用いられる略語や省略形まで含め,それらの英語と日本語表記を網羅するシノニムテーブルを構築した。次にこれらを用いて,化学物質と生体分子の相互作用を数少ない相互作用様式とともに三項関係によって記述した新たな分子薬理学データベースを構築した。

2. 方法

JAPIC医療用医薬品集および日本医薬品一般名称(JAN)より抽出した商品名,製剤名,有効成分名を基本にして,欧米で用いられる名称および表記をATC,KEGG,MeSHより追加した。こうして集めた日本語および英語の名称について,有効成分が同一なものの集合を同義語として整理した。漢方製剤は有効成分が明らかな場合以外は除外した。MeSHに収録されている場合は,Descriptor表記を代表記述子とした。次に,ライフサイエンス辞書(LSD)[6]に収録されている英和対訳の物質名をMeSHツリーと関連づけ,生体分子名の同義語テーブルを構築した。

作用点の定義には,国内外の薬理学テキストと独自に収集したPubMedコーパスの解析結果を参考にした。これらテキスト(500 MB)に出現する医薬品および生体分子にタグを付与し,続いてタグ同士の共起頻度をperlスクリプトで解析することによって相互作用と考えられるペアを抽出した。この化合物と標的生体分子との相互作用について,様式を表現する阻害,活性化などの少数の記述子を用いて三項関係として記述した。標的分子が明らかでない場合は,細胞ないし組織レベルにおいて報告されている知識を,別の関係テーブルにおいて同様にMeSH用語と少数の作用様式を用いて整理した。以上のデータはFileMaker Proを用いたリレーショナルデータベースとした。

3. 結果

図1 医薬品名の同義語テーブル

収集した医薬品名称(商品名,製剤名,有効成分名)は約18,000種類(うち,日本語は約6,000)であった。これらを有効成分によって整理し,2,842種類の有効成分リストを作成した(図1)。このうちMeSHには2,545種類,ATCには2,599種類が収録されており,どちらにも収録されていない成分はなかった。次に,各有効成分について,一般的に行われている薬効分類を適用し,化合物群によるグループ化を行った。一方,LSDから抽出した生体分子名は英和対訳として19,595種類であった。これらをMeSHツリーを参考にして同義語テーブルとして整理した結果,4,272種類の標的候補分子リストを作成した。

上述の医薬品名および標的候補分子名が薬理学教科書およびPubMedコーパスのテキスト中で共起する頻度をカウントした結果,ほとんどの医薬品について共起頻度の上位ペアに主作用点が浮かび上がった(図2)。そこで,このデータを参考にして,個々の薬物について薬理作用点である標的分子を特定した。この際,同一の薬効分類に属する医薬品について,同一の作用点を有するかのチェックを行った。

図2 化合物と生体分子の関係抽出

各々の相互作用様式については,標的分子が受容体,酵素,膜輸送タンパク質といったタンパク質の場合には,その結合様式が非共有結合であっても共有結合であっても,結果としてそのタンパク質の機能に与える影響を阻害,活性化など少数の記述子を用いて表現した(図3)。

図3 分子作用点テーブル

また,転写調節に影響する薬物の場合は,主たる標的である結合タンパク質とともに,結果的に発現が調節される遺伝子のうち,主たる薬効の発揮に関与していると考えられる標的タンパク質についても発現上昇,発現抑制などの記述子とともに収録した。同様に,ホルモン補充療法などの場合,類似した薬理作用をもつ生体成分の補充であることを示すとともに,生体成分が作用する標的を記述した。これらの収録にあたっては,医薬品名にMeSHおよびLSDでのコード番号とともに,生体分子にはMeSH Descriptor IDおよびRefSeq遺伝子名を付与し,他データベースへの互換性と拡張性を持たせた。

以上の結果,全体の70%に相当する薬物について,計500種類の標的分子との特異的な相互作用を収録した。分子レベルの作用が不明瞭で,細胞ないし臓器レベルでの効果が知られている残りの薬物については,それらを別テーブルにおいてMeSH用語と相互作用記述子を用いて収録した。

4. 考察

約2,800種類の医薬品有効成分のうち,70%に分子作用点が記述でき,その作用点が500種類であったという結果は,過去に調べられた医薬品の標的についての調査とよく一致している[7,8]。標的分子が1つ以上に及ぶ薬物は約1,000種類と多く,薬効を考える上で必要な知識を網羅していると考えられる。しかしながら,MeSH収録語を中心とした記述では一部において表現できないサブタイプが存在することが明らかになった。また,今回は相互作用を定性的にのみ表現したが,実際のテキスト解析に応用する場合を想定すると,複数の作用点に異なる親和性をもって作用する場合などで定量的な尺度を導入する必要性が感じられた。

医薬品の同義語関係を整理するとともに分子作用点を正規化した本データベースは,将来的に医薬品の適応症や副作用,あるいは代謝パスウェイ等のデータベースと組み合わせることによって有害作用の予測や因果関係の解析に有用な資源となると考えられる。

参考文献

[1] KEGG生命システム情報統合データベース.http://www.kegg.jp/.京都大学化学研究所.
[2] OmicSpace.http://omicspace.riken.jp.理化学研究所.
[3] 金子周司(辻本豪三,田中利男編).創薬統合データベース.21世紀の創薬科学.共立出版, 1998: 141.
[4] Wishart DS et al.DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nuc. Acids Res. 2006; 34: D668.
[5] 吉川澄美,松村和美,小長谷明彦.薬機能オントロジー:分子機能と生体現象の関連化.人工知能学会第5回セマンティックウェッブとオントロジー研究会抄録集, 2004.3.
[6] ライフサイエンス辞書.http://lsd.pharm.kyoto-u.ac.jp/.京都大学薬学研究科.: 1960.
[7] Drews J.Drug discvery: a historical perspective. Science 2000; 287: 1960.
[8] Hopkins AL and Groom CR.The druggable genome. Nature Rev. Drug Discov. 2002; 1: 727.

医学用語シソーラスに基づく効率的医療情報検索システムの開発

金子 周司[1] 鵜川 義弘[2] 大武 博[3] 河本 健[4] 竹内 浩昭[5] 竹腰 正隆[6]
天野 博夫[1] 藤田 信之[7]

京都大学大学院薬学研究科[1] 宮城教育大学[2] 京都府立医科大学[3]
広島大学医歯薬総合研究科[4] 静岡大学理学部[5] 東海大学医学部[6]
製品評価技術基盤機構[7]

Development of medical portal system based on the thesaurus and collocation analysis of biomedical terms

Kaneko Shuji[1] Ugawa Yoshihiro[2] Ohtake Hiroshi[3] Kawamoto Takeshi[4]
Takeuchi Hiroaki[5] Takekoshi Masataka[6] Amano Hiroo[1]
Fujita Nobuyuki[7]

Kyoto Univ. Grad. Sch. Pharm. Sci.[1] Miyagi Univ. Edu.[2] Kyoto Pref. Univ. Med.[3]
Hiroshima Univ. Grad. Sch. Biomed. Sci.[4] Shizuoka Univ. Fac. Sci.[5] Tokai Univ. Sch. Med.[6]
NITE[7]

The internet has become a commonly used and popular form of media by which health care information can be searched and retrieved not only by general populace but also by medical students and professionals. Accordingly, an increase in the number of Japanese documents easily available through the Internet has caused the relative avoidance of original scientific papers and databases written in English. To develop a translational portal site that enables the use of Japanese for medical infomation retrieval, we have constructed an English-Japanese thesaurus containing 160,000 terms collected from biomedical literatures. We have assigned individual terms to the MeSH descriptors (25,000 English-Japanese pairs), and analyzed the top-30 collocations of the terms in medical literature (total 1.4 million pairs). The collocation data are available for associative search, equipped with a tree-style thesaurus in the online Life Science Dictionary (LSD). The free LSD portal site enables the use of Japanese terms for information searches linked to Entrez-PubMed and Google sites. In addition, we have developed a new dictionary for Mac OS X 10.5 that can be used in the Safari browser as a one-touch dictionary. The development of new portal systems and dictionaries will be useful in medical education and research activities.

Keywords: Thesaurus Tree, Synonyms, Dictionary, Associative Search

1. はじめに

インターネットを介して得られる医療健康情報は,一般市民だけでなく医療従事者および医療系学生にも影響を及ぼしつつある。特に,日本語で記述された解説記事などの増加は,相対的に英語で記述された原著論文など,科学的に評価の高い一次資料が利用されなくなる状況を招いている。しかし,もし英語のリソースを日本語で検索できるポータルが提供されれば,利便性と有益な波及効果が期待できる。そこで本研究では,インターネットで公表される広範な医学関連研究の成果を,日本人が検索あるいは理解しやすくすることを目的として,過去10年以上にわたって構築してきたライフサイエンス辞書[1,3]の資源を利用するシソーラスの制作と日本語ポータルの研究開発を計画した。

本研究のゴールは,第一に医学研究論文やデータベースを日本語で検索する場合に,表記のゆれや翻訳を実装するとともに,入力したキーワードと密接に関連する別のキーワードを同時に提示することによって情報検索を容易にする連想検索サーバを開発することとした。また第二に,検索結果として表示される英語ページにおいて,利用者が求める箇所でオンデマンドに専門用語の対訳および解説を表示して,利用者の理解を助ける情報ポータルの試作を行った。英語で書かれた医学情報のあらゆるWebページを日本語で検索して内容を理解できるサーバを無料で公開することで,医療および医学研究の成果を広く社会に提供する実用的なインターフェースとして幅広い利用や応用が見込める。

2. 方法

2.1 シノニム辞書の制作

医療情報の理解に必要と考えられる専門用語の異表記を統一するための統制語としては,後で情報検索に利用することを考えた結果,まずはアメリカNational Library of Medicineが構築しているMedical Subject Headings (MeSH)に準拠することにした。そこでMeSH 2008(2007年11月版)より解剖部位(Tree A01-A17),生物名(Tree B01-B08),病名および症候名(Tree C01-C23および精神疾患F03),生体分子および医薬品名(Tree D01-D27およびSupplemental Concepts),方法および尺度(Tree E01-E07),学問領域や現象(Tree G01-G14)に帰属する専門用語から,上記のカテゴリに帰属できるDescriptor 21,684語とLSDに同一の見出し語が収録されていたSupplemental Concepts 2,668語を合わせた計24,352語を統制語として採用した[2]。これを元に,LSDとMeSHのすり合わせ作業を行い,シノニム辞書を制作した。また,一部の統制語においては,MeSHに準拠せず,出現頻度

の高い語句を採用した。

2.2 共起する統制語による関連概念データの制作

PubMed抄録を収集した文献コーパスに対して,シノニム辞書を適用して統制語によるタグ付けを行うPerlスクリプトを開発した。統制語タグが同一抄録中で共起する頻度を解析し,各用語について出現頻度,共起する他の統制語およびその共起頻度を得た。得られたデータが専門的に見て妥当な関連性を表すかどうかを,複数名の研究者による目視によって検討した。この評価に基づいて,検索キーワードの取捨選択を行い,最適化を試みた。

2.3 関連概念を提示する情報検索エンジンの開発

シソーラスと共起解析データをオンライン版ライフサイエンス辞書WebLSDに実装することによって,日本語および英語のいずれによっても表記のゆれを吸収して統制語による情報検索を可能にするポータルシステムをPerl cgiにて開発した。

2.4 日本語訳を表示する辞書ツールの開発

ウェッブブラウザで表示されるゲノム情報などの英語ページにおいて,可能な限り簡単な操作で専門用語を辞書引きできるツールを開発するため,Mac OS X 10.5においてシステム標準で利用できる辞書.appでの試作と検証を試みた。この辞書.appではブラウザであるSafariからショートカットで複合語レベルでの辞書検索が実現できる。また,辞書を制作するためのアプリケーションやテキスト使用がAppleにおいて公開されている。

3. 結果および考察

3.1 シノニム辞書の制作

2008年6月時点で,表1に示すカテゴリのMeSH Descriptor および Supplemental Concepts (SC) に帰属する統制語2.5万語の96%を日本語化し,英語表記と日本語表記を併記できるようにした。その上で,延べ約16万語の英語および日本語で記述されるLSD収録語およびMeSH用語を統制語に集約することで,対訳シソーラスとシノニム辞書を制作した。このデータから,16万語の同義語のうち,LSD収録の英語と日本語,および新たに加えたMeSH英語が,それぞれほぼ3分の1ずつの割合を占めることがわかる。生体分子などの物質名,特に海外での医薬品商品名や化学一般名などの異表記を非常に数多く含む物質カテゴリにおいては,MeSHに由来する名称が半数に及び,これら新しく加えた用語によって欧米の文書に対する網羅性が高まったことが期待できる。

表1 ライフサイエンス辞書のシソーラス化(概要)

MeSH Tree	カテゴリー	統制語数(a)	シノニム数(b)	平均異表記数(b/a)	LSD英語		LSD日本語		MeSH物質	
A	解剖部位	1,502	7,507	4.9	3,309	47%	3,080	44%	653	9%
B	生物名	3,478	16,499	4.7	5,157	31%	8,870	42%	4,473	27%
C+F03	病名・症候名	4,533	38,621	8.2	8,315	35%	11,495	44%	5,807	23%
D	物質名	11,260	91,090	8.1	20,571	33%	25,479	29%	44,841	49%
	(うち医薬品)	3,969	42,367	11.9	7,749	19%	11,072	26%	23,566	56%
E	方法, 尺度	2,165	11,071	5.1	3,663	33%	4,676	42%	2,732	25%
G	知識, 現象	1,570	7,332	4.6	2,734	37%	3,024	41%	1,574	21%
計		24,352	159,835	6.6	44,753	28%	55,603	34%	60,078	38%

百分率はシノニム数(b)に対する割合を表す。

しかしながら一方で,LSDに収録されながらMeSHと照合できないため統制語に帰属されない用語が英語で1万語以上も存在することが明らかになった。特に,病名・症候名や解剖部位名においては,国内で用いられている標準病名マスターや国際的な有害事象報告のための統制語であるMedDRAにも収録されながらMeSHに帰属できない用語が数多く残された。また,医薬品としては国内医薬品において収録されていない用語が多数存在した。今後は,これらの語句を帰属させるためにツリーを拡張していく必要が示された。

3.2 共起する統制語による関連概念データの制作

PubMedより代表的な学術誌に掲載された10年分の論文抄録(600 Mバイト)をコーパスとして収集し,シノニム辞書によってテキスト中に最長一致で統制語のXMLタグを施した。このタグ付けテキストの内容をブラウザで確認しながら,曖昧性の排除と統制語の最適化を行った(図1)。この過程において,テキストでの一致のみによって統制語への変換を行う場合,曖昧性を排除するために多義性のある略語や商品名等,一部のシノニムをタグ付け辞書から除外する必要が生じた(約200語)。また,「ヒト human」,「病気 disease」,「酸 acid」等のように,非常に大きな概念は関連するキーワードとして不必要あるいは不適切と考えられたため,それら(約360語)もタグ付けから除外した。

図1 英語論文抄録にタグ付けを施したXMLデータをブラウザで表示した例

テキスト中に出現する専門用語をすべて統制語に自動変換するPerlスクリプトを用いてXMLデータを作成した。日本語の統制語見出しを,物質や医薬品名は青色で,病名は赤色で,方法や尺度は緑色で表すことによって抄録で述べている内容に関連するキーワードの関係や,統制語の妥当性を判定できるようにした。この抄録の場合,「ベンゾジアゼピン」と「股関節骨折」の関係を述べている論文であることが一見してわかる。しかし「recipient」を「移植」という統制語に翻訳した箇所は誤りであるため,このような対応関係は解析辞書から除外する措置をとることによって最適化を行った。

次に,同一抄録中で共起する統制語のペアを収集することによって計100万対以上の統制語の共起頻度を求め,出現した2万語の統制語ごとに上位30対までの共起概念データを得た。ここで解析に用いるコーパスによって得られる共起概念のリストは大きく異なった。例えば,1.3 Gバイトの臨床研究抄録を用いた解析では,ある薬物と共起する概念はほとんど医薬品

名で占められ,標的となる生体分子や作用メカニズムを示唆するキーワードが得られない等,必ずしもコーパスが大きいからと言ってデータが適切にならない場合も見いだされた。

本研究は医療系学部あるいは大学院に所属する学生による教育現場での利用を想定していたため,上述した代表的学術誌にこの10年間に発表された先端的研究成果を記述する広範な学術分野の論文抄録に限ることとした。試行錯誤の末,主観的に見てほぼ全ての用語カテゴリーにわたってバランス良い共起結果を得ることができたと考えている。しかし今後,コーパス母集団を変えることでさらに専門家の知識を反映するような最適化を試みる必要があると思われる。

図2 WebLSDに実装したシソーラスと共起概念による医療情報ポータルの使用例
(左上)「とうにょう」と入力したときに表示される和英辞書で「糖尿病」シソーラスをクリック(右上)用語ツリーから下位概念である「2型糖尿病をクリック」(左下)2型糖尿病の共起概念リストから「遺伝子多型 Genetic polymorphism」をクリック(右下)PubMedに「type 2 diabetes mellitus」と「genetic polymorphism」の2つのMeSHが渡されて,ヒットする文献リストが表示される

3.3 関連概念を提示する連想検索エンジンの開発

このようにして得た共起概念データをシソーラスのツリー表やシノニム表示と組み合わせることによって,検索語として入力した日本語あるいは英語を自動的に統制語に直して表示するだけでなく,ツリーによって上位や下位の概念を探索できるようにしたり,関連性の高い共起概念を表示することで既存ポータルに適切なキーワード対を検索語として渡したりするためデータをXML形式で制作した。

これらデータをまずウェッブブラウザで検索可能にするため,公開しているWebLSDのサブセットとして,英和・和英対訳辞書と一体で使うことができるようなcgiを制作し,2008年6月より公開した。

このWebLSDに実装したシソーラスを用いることによって,任意に入力する日本語の検索語が英語に訳されるだけでなく,MeSHに準拠した統制語について,シノニム,ツリー,共起概念が表示される(図2)。ツリーでは表示している統制語が赤字で表示され,その上位と下位に位置する概念へクリックで自由に移動することができる。共起概念は日本語と英語で最大30種類がそれぞれの統制語ごとに表示され,日本語をクリックした際には選んだ用語と統制語との組み合わせでGoogleへ検索キーワードが渡される。また,英語のリンクからはEntrez-PubMedにキーワード対が渡されるようにしてある。基本的にはURLを明示できる検索エンジンやデータベースに対して,このインターフェースを介してデータを渡すようにカスタマイズすることは容易にできるため,汎用性や応用性にも優れている。

3.4 日本語訳を表示する辞書ツールの開発

Mac OS X 10.5には標準で辞書ツールである辞書.appが付属している。この辞書.appは日本語にも対応しており,キーワード入力に応じて結果を表示するincremental searchを可能にした特徴を持っている。また,辞書.appはMac OS X 標準ウェッブブラウザであるSafariからショートカットキー(Command + Control + D)によって呼び起こすことができる。さらにこの時,カーソルが置かれている単語の前後を最長一致で判定し,最もその場所にふさわしい複合語を選び出して表示する他の辞書には見られない機能を有している。Apple社は辞書.appに対応する辞書を制作するための技術資料を公開しているため,今回,この辞書.appを用いる辞書を制作した。

その結果,WebLSDで実装したシソーラスとほぼ同様の機能を有するスタンドアロン辞書を制作することができた(図3)。辞書.appは検索語の途中でも先読みでキーワードを表示するため,前方一致するキーワードリストを見ながら,適切な用語のシソーラスを見ることができる。シソーラス内での操作はほぼWebLSDと同様であり,ツリーの上下移動や共起概念からの外部リンクを装備することができる。また,このようにして制作した辞書はSafariに表示されたhtmlページのカーソル位置からショートカットキーで呼び出すことができるため,英和の用語検索が容易に行える。今後,さらに辞書.appおよびSafariの連携が簡単かつ高機能になることを期待したい。

図3 Mac OS X辞書への実装
Safariで電子ジャーナルを検索した際にショートカットキーで辞書.appを起動した画面

4. おわりに

以上のように,本研究では当初の計画で予想した以上に有用なポータルを開発することができた。今後さらにデータの最適化を計り,公開ポータルとしての利便性を向上させる予定である。また,提示する共起概念の視覚的な表示技術についても検討していきたい。しかし,教育における本シソーラスの利用経験はまだ浅いため,将来的にこれらを医療情報教育に活用し,客観的な評価も進めたいと考えている。

5. 謝辞

本研究は(財)電気通信普及財団研究調査助成(平成18年度),厚生労働省科学研究費(平成18-20年度)および(独)日本学術振興会科学研究費研究成果公開促進費(平成17-19年度,177002)の研究助成を受けて行われた。辞書.appについては中村浩之氏から制作するきっかけとなる示唆をいただいた。ここに記して感謝の意を表したい。

参考文献

[1] 金子周司,鵜川義弘,大武博,河本健,竹内浩昭,竹腰正隆,藤田信之.ライフサイエンス辞書2の制作と公開.コンピュータサイエンス,Vol. 2, No. 2, 135-142, 1995.
[2] 金子周司,藤田信之.文献情報の解析に基づく対訳シソーラスの評価.医療情報学,Vol. 25, No. 6, 475-483, 2005.
[3] 金子周司.ライフサイエンス辞書とは.情報管理,Vol. 49, No. 1, 24-35, 2006.
[4] ライフサイエンス辞書.http://lsd-project.jp/.

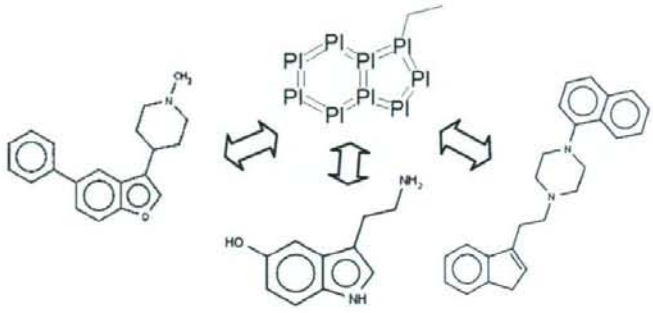
JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

Substructure Mining of GPCR Ligands Reveals Activity-Class Specific Functional Groups in an Unbiased Manner

Eelke van der Horst, Yasushi Okuno, Andreas Bender, and Adriaan P. IJzerman

J. Chem. Inf. Model., **2009**, 49 (2), 348-360• DOI: 10.1021/ci8003896 • Publication Date (Web): 03 February 2009



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Substructure Mining of GPCR Ligands Reveals Activity-Class Specific Functional Groups in an Unbiased Manner

Eelke van der Horst,[†] Yasushi Okuno,[‡] Andreas Bender,[†] and Adriaan P. IJzerman*[,†]

Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC Leiden, The Netherlands, and Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University, Japan

Received October 21, 2008

In this study, we conducted frequent substructure mining to identify structural features that discriminate between ligands that do bind to G protein-coupled receptors (GPCRs) and those that do not. In most cases, particular chemical representations resulted in the most significant substructures. Substructures found to be characteristic for the background control set reflected reactions that may have been used to construct this library, e.g. for the ChemBridge DIVERSet library employed these are ester and carboxamide moieties. Alkane amine substructures were identified as most important for GPCR ligands, e.g. the butylamine substructure, often linked to an aromatic system. Hierarchical analysis of targeted GPCRs revealed well-known motives and new substructural features. One example is the imidazole-like substructure common for the histamine binding receptor ligands. Another example is the planar ring system consisting of a fused five- and six-membered ring (indole-like substucture) common for the serotonin receptor ligands.

INTRODUCTION

Chemical structure mining has a long tradition in the prediction of molecular properties. Methods for analyzing the structural features of molecules can be broadly divided into two categories: methods that focus on predefined structural parts (fragments) and methods that consider the complete set of possible substructures of a molecule.[53] Methods of the first category apply a set of fragmentation rules to partition the molecular structure into discrete fragments, which are then analyzed. Examples of such fragments are ring systems, linkers, and side chains,[1−5] synthetic building blocks,[6,7] or algorithmically defined molecular fingerprints.[8,9] Analysis of fragment frequencies has proven useful for the description and comparison of molecular databases and for the identification of 'chemical clichés'.[10,11] For instance, unexplored parts of chemical space, with only a few fragments, become apparent as well as the preferences of chemists for certain reaction types or starting materials, yielding a more densely populated chemical space. Analyzing the co-occurrence of fragments may further yield valuable information, i.e. on pairs that seem to avoid each other or pairs that constitute a common template. Analyzing fragment occurrences may also aid the design of new ligands in (chemical) fragment-based drug discovery[12] and is a prerequisite for similarity searching,[13] an approach in which predefined structural parts are utilized to construct molecular fingerprints. A fingerprint is a reduced representation of the molecule that holds information on the presence or absence of certain features.[13] Features included in the fingerprint may be aforementioned (discrete) fragments, such as rings and functional groups (so-called structural keys, e.g. MDL keys[14]), but also algorithmically defined, such as arbitrary structural elements of fixed size, of which the circular fingerprint has gained some popularity.[15] Since predefined fragmentation rules are dependent on the choices of the chemist, analyses and predictive models are inherently biased.

Complementary to analyses using predefined structural fragments as well as fingerprints are methods that consider all possible substructures that are found in the 2D structure of a molecule. These substructure-based methods thus avoid the bias that is intrinsic to the use of predefined fragments. However, they come at the price of computational expense. In a simple structure as for the amino acid alanine without explicit hydrogens, the number of substructures amounts to 20 already. Adding two methyl groups, i.e. the amino acid valine without explicit hydrogens, will yield 39 substructures. Because of the exponential growth of substructure count with increasing molecule size, most substructure methods seek ways to limit the number of substructures to be evaluated. Batista et al. described a method that uses repeated random fragmentation of molecules to generate profiles that served as a measure of molecular similarity[16] and later applied this to database screening.[17] Although this work represents substructure analysis in an unbiased manner, the success of the method depends on the choice of parameters (iterations, number of bonds cut). Besides that, the set of evaluated substructures is not complete, i.e. the evaluation of a substructure depends on chance and not on occurrence in the molecules. Two other methods that are substructure-based are maximal common substructure analysis and frequent substructure mining. Maximal common substructure analysis finds the largest connected substructure that a certain number of molecules have in common.[18,19] It is used for similarity

* Corresponding author phone: +31 71 5274460; fax: +31 71 5274565; e-mail: ijzerman@lacdr.leidenuniv.nl.
[†] Leiden University.
[‡] Kyoto University.

10.1021/ci8003896 CCC: $40.75
Published on Web 02/03/2009

and SAR analysis, for instance as implemented in commercial tools such as Pipeline Pilot (Scitegic) and ClassPharmer (Simulations Plus).[20,21] Frequent substructure mining finds the most common substructures in one or more sets of molecules by considering all substructures that occur in the molecules. It uses a minimum-frequency constraint to control the amount of substructures that are evaluated. It is an application of frequent subgraph mining, which finds all frequently occurring connection patterns from a set of graphs. Recent advances in graph-mining algorithms, together with the steady growth of computing power, have made it possible to mine data sets as large as 200,000 molecules on a standard PC.[22] Frequent substructure 'miners' have been successfully applied for prediction of CNS activity, bioactivity, and toxicity.[23−25] The SUBSTRUCT program developed by Engkvist et al. distinguished CNS active compounds with approximately 80% accuracy.[23] However, their approach was bound to a maximum size of the generated substructures, which was between 1 and 4 atoms by default. In contrast, Borgelt et al.[24] identified substructures that model the activity classes for HIV-1 antiviral screening data. Similarly, Kazius et al. applied the frequent subgraph miner Gaston[26,27] to mutagenic compounds and extracted a decision list of six discriminative structural features associated with mutagenicity.[28] For this, the authors adopted several different types of chemical representation. So-called elaborate chemical representation adds extra information to a molecule, for instance by adding extra labels to atoms or by replacing certain atoms with wildcards (abstractions). The authors obtained the most significant results when elaborate chemical representation was used. Similarly, others also reported improved findings when using, for instance, abstractions for rings and chains (reduced graphs).[29,30]

In the context of GPCR (G Protein-Coupled Receptor) ligands, the most important source of current medicines, only methods that analyze discrete fragments have been described. For instance, a property-based scoring scheme was constructed for the classification of GPCR ligands, intended for the creation of focused libraries.[31] Similarly, Schnur et al. identified frameworks or substructure classes that are common for families of ligands. In the context of GPCR ligands, these were defined as privileged structures albeit that these frameworks were not selective for GPCRs.[32]

While previous studies were limited to analysis of predefined fragments, in this study, we will use a complete method (i.e., frequent substructure mining) to analyze the structural features of GPCR ligands. This method represents an unbiased as well as an exhaustive way to mine information contained in chemical data sets of GPCR ligands. We build upon the approach described by Kazius et al.[28] to find frequently occurring substructures that are discriminative for GPCR ligands. This is accomplished by comparing the ligands against a control group and analyzing the frequencies of all possible substructures that occur in the sets. To include additional chemical details, both normal and elaborate chemical representations (atom and bond type abstractions with atom labels) were used. However, abstractions for molecular parts, such as special types for rings or chains, were omitted since these depend on the choice of the chemist, thereby introducing a bias. In addition, with reduced graph representations, information such as bond distance or substituent positions is lost, which led us to believe that the choice of the current algorithm is appropriate for the work performed here. To derive the significant features common to specific groups of ligands, we conducted two additional experiments on subsets of the original sets. For the first experiment, subsets were based on the presence of the previously found most-significant substructure. For the second experiment, ligands were grouped into subsets according to the hierarchical classification of their target GPCRs. In addition to the work of Kazius et al.,[28] we also analyzed which substructures are rarely found in the GPCR ligands compared to the control group. This type of analysis would be less useful for prediction of mutagenicity but has added value for prediction of receptor binding. In the latter case, there will be substructures that contribute not only to binding but also to lowering this possibility (e.g., steric hindrance, unfavorable pharmacophoric features). From this analysis, we established a comprehensive list of favorable and unfavorable features of this important ligand class.

MATERIALS AND METHODS

Data Sets. GPCR ligands were collected from two publicly available online resources: the GPCR-ligand database (GPCR-Ligand Database or GLIDA) from the University of Kyoto[33] and a database for human GPCRs and their ligands (hGPCR-lig).[34] The set from GLIDA consisted of 22,122 ligands for human, mouse, and rat receptors, collected from various public data sources, such as PubChem,[35] K_i Database,[36] and scientific literature. The 17,908 GPCR ligands from the hGPCR-lig database are taken from the scientific literature and the MDL Drug Data Reports (MDDR). Although the sets partly overlap, the first one is more illustrative for published research from academia, while the second is more representative for patented drugs recently launched or under development, i.e. commercial drugs, preclinical candidates, etc. These two sets were compared against a control set, denoted as the background set. For this, 15,993 compounds from ChemBridge's DIVERSet screening collection were used.[37] This set was meant to provide a suitable contrast to the GPCR ligands since we were interested in how GPCR ligands differ from a diverse collection of small molecules. In GLIDA, ligands are grouped according to the classification of their targets. Targets follow the pharmacological classification of GPCRs as used by the International Union of Pharmacology (IUPHAR).[38,39] Targets are arranged into a hierarchy of subfamilies, families, and classes, as used by the GPCRDB information system.[40] In GLIDA, some receptor classes are missing due to the low number of known ligands, e.g. trace amines in amine-binding GPCRs. Each ligand-target pair is annotated with an activity type, namely full agonist, partial agonist, agonist, antagonist, or inverse agonist. Since the annotation of GLIDA is not entirely completed yet, we only used 'agonist' (for full, partial, and agonist) and 'antagonist' (for inverse agonist and antagonist). Classification of compounds as active depends on the origin of the compounds. A reported affinity in one of the source databases classified a compound as active, independent of the reported binding affinity. The sets were cleaned and standardized using Scitegic's Pipeline Pilot 6.1.5.0 Student Edition.[20] Salts, counter-ions, and other small fragments associated with the molecules were removed, and zwitterions were neutralized. Charge and stereochemical information was discarded, and bonded hydrogen atoms were omitted from

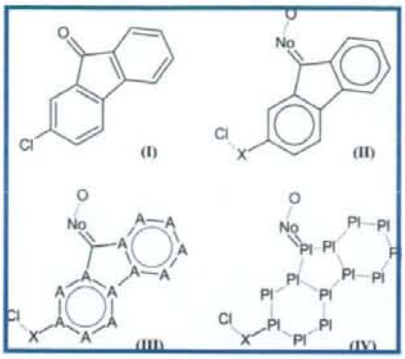


Figure 1. Example molecule in normal (**I**) and two elaborate chemical representations: one for aromatic bonds (**II**), one for aromatic atoms and bonds (**III**), and one for planar ring systems (**IV**). In the normal representation (**I**), aromatic bonds are represented as alternating single and double bonds, whereas in the first elaborate representation (**II**), a special type for aromatic bonds is used. In addition to this representation, the representation in (**III**) adds a special type for aromatic atoms (A). Since this example molecule has a planar ring system, all atoms that constitute the system are denoted as 'Pl' in the planar representation (**IV**). In both elaborate chemical representations, wildcards are used for heteroatoms ('No') and for halogens ('X') with a label attached specifying the actual atom-type.

the representation. After that, ChemAxon's standardizer[41] was used (for consistency with existing databases) to convert the structures into a uniform representation and to filter out duplicates. Only structures with molecular weight below 1000 Da were used. The final GPCR ligands sets consisted of 21,619 compounds for GLIDA, 16,509 compounds for hGPCR-lig, and 15,983 compounds for the background set. In some cases, analysis of large data sets using elaborate representation (see below) proved to be difficult since physical limits of system resources (maximum file size) were reached. In these cases, the experiment was continued using a sampled set of 5k ligands. These 'sampled sets' were constructed using Pipeline Pilot's Random Percent Filter.[20]

Chemical Representation. Molecular structures are represented as labeled graphs. Hydrogen atoms were excluded from representation. Four types of chemical representation were used: the initial chemical structure representation with the atom and bond types unchanged and three 'Elaborate' Chemical Representations (ECRs).[28] Figure 1 offers an example that accompanies the following description of the representations. Elaborate representation is a method to include extra information about the molecule by using abstractions, translations, and/or extra labels. The first elaborate representation includes a special bond type for aromatic bonds. In addition, the second one has a special type for aromatic atoms. The third representation offers a special type for planar ring systems, which has been successfully applied previously to predict the mutagenicity of compounds.[28] In elaborate chemical representation, aliphatic Nitrogen, Oxygen, and Sulfur atoms were represented as an aliphatic heteroatom by replacement with the symbol No. An extra label was attached to N and O to indicate the type and number of bound hydrogens, 'Ze' (zero) for no bonded hydrogens, 'On' for one bonded hydrogen, and 'Tw' for two bonded hydrogens. The halogen atoms, Cl, Br, I, and F, were replaced by X, and an extra label was attached to indicate their type. Note that the above-mentioned representation differs from Kazius et al.[28] in that only one heteroatom type was used and not two types for 'small heteroatom' (No) and 'large heteroatom' (Ps). In addition, fluor (F) was included as part of the halogen abstraction. Aromatic atoms and bonds were detected with basic aromaticity. Figure 1 has an example of a molecular structure in normal and chemical representation. The use of alternate representations may cause the same graph to appear multiple times. The aim of abstractions for atom and bond types is to raise the occurrence of similar substructures above the support threshold. Individually, these substructures might go undetected; however, the occurrence of their common representation sums the individual frequencies.

Frequent Substructure Mining. The frequent subgraph-miner Gaston was used to find all frequently occurring substructures in the data sets.[26,27] Frequent subgraph miners such as Gaston iterate over all molecules, extracting all possible substructures per molecule. Current subgraph miners utilize several approaches to keep the number of found substructures to a minimum. One reason is that a larger substructure can never occur more frequently than the smaller substructures it consists of. This allows numerous substructures to be pruned before being considered. Compared to other algorithms, Gaston is more efficient since computationally expensive operations take place in the last steps, when a large number of possible substructures has already been discarded. For a quantitative comparison of Gaston with other frequent subgraph miners, see ref 22. The importance of a substructure was determined by comparing its frequency against the frequency of occurrence in the control set. The most revealing substructures are those that occur frequently in one set and not in the other. As a measure of the importance of a substructure, the significance of association with one of the sets was determined by calculating the p-value of the finding. The p-value as used in this study is defined on page 3 of the Supporting Information of ref 42. It is the probability to find a statistical association with one of the two groups based on chance alone. On the assumption of a binomial distribution, it was calculated based on the number of ligands versus control group that were detected using that substructure. While this measure makes assumptions such as to the underlying distribution of features in each database, we still found it to be useful also in the ranking scheme described here. Using the p-value, the lists of frequently found substructures were ordered according to significance with the most-discriminating substructures at the top. The substructure with the lowest p-value was considered the most significant finding. The p-values of substructures are the same if they have the same absolute frequency. When two substructures had the same p-value and one substructure was a substructure of the other substructure, only the larger substructure was kept. On average, this was the case in 25% of all substructures. In case of substructures with equal p-values that were not substructures of each other, the larger substructures had preference over smaller ones in the list ordering. For example, if alanine and valine would be substructures with equal occurrences, and hence equal p-values, only the valine substructure would be kept in the list since alanine is a substructure of valine. In the case of leucine instead of alanine, both substructures would be kept since neither of the two is a substructure of the other. Another important parameter in frequent subgraph mining is the

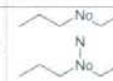
minimum support value, which is the relative number of molecules a substructure should occur in to be detected by the algorithm. Lowering the minimum support will result in finding an equal or higher number of substructures. A higher number of substructures increases the chance of finding a substructure that is more significant. However, there is a balance between minimum support and p-value of the most significant substructure. This will be illustrated with the following example of two sets of 100 compounds each. Presume that the most significant substructure found at a support threshold of 30 compounds occurred in 60 active and 20 control compounds. The p-value for this substructure is 5.32e-09, which is the chance of finding this substructure based on chance alone. Lowering the minimum support from 30 to 20 means that a new set of substructures is added to the already generated set. In theory, the most significant substructure that could be added with the new set has an occurrence of 29 active and 0 control compounds. The p-value of this theoretical substructure is 1.73e-10, which is more significant than the actual found substructure (with a p-value of 5.32e-09). Therefore, the experiment should be repeated at a lower minimum support of e.g., 20 to examine whether this theoretical substructure does actually exist. The experiment is completed if it results in the same substructure as found in the first run. This is because a more significant substructure cannot be found by lowering the support, i.e. the best theoretical substructure (19 active, 0 control) has a p-value of 7.42e-07. When another, more significant substructure is found at a lower support, the process is repeated until no theoretical substructure can be found that is more significant. Concluding, the minimum support value was chosen by iteratively lowering it per run until no better (more significant) substructures could be found, resulting in practice in support values between 10% and 30% for the data sets used here.

Software. For translating the molecules into elaborate representation, for partitioning the graph sets, and for the p-value calculations simple awk and bash scripts were used. These scripts and the substructure mining were run on Scientific Linux. (Sub)structures were edited and visualized using (MS Windows based applications) Pipeline Pilot 6.1.5.0 Student Edition and MDL ISIS/Draw 2.5.[20,43]

RESULTS AND DISCUSSION

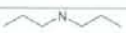
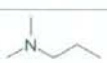
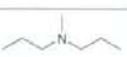
Mining of Characteristic Substructures of GPCR Ligands. Frequent subgraph mining was first applied to each individual set, both hGPCR-lig and GLIDA, for a broad analysis of structural features in GPCR ligands. Analysis of the substructure distributions revealed the best discriminating substructure for each of the four elaborate representations (see Materials and Methods). The substructure originating from the 'normal' representation performed best in discriminating GPCR ligands (both from hGPCR-lig and from GLIDA) from background compounds, i.e. the ChemBridge DIVERSet library. The statistics for all representations are summarized in Table 1, demonstrating that within each of the four representations highly significant substructures are occurring. The largest substructure was found in the 'aromatic atoms and bonds' representation and suggests a symmetrical organization of lipophilicity (through aliphatic carbon atoms) around a heteroatom, which was specified as nitrogen for GLIDA. This chemistry implies that at physiological pH the nitrogen heteroatom is likely to be protonated and charged, in line with the notion that many GPCR ligands interact with biogenic amine receptors. Strader et al.,[44] in one of the first mutagenesis studies on GPCRs, identified a negatively charged aspartic acid residue in transmembrane domain 3 of the β-adrenoceptor to form a salt bridge with the ligands' protonated amino group. A similar, though one atom smaller, substructure is found in the 'normal' representation. Consequently, this substructure is also at the top of the list of most significant substructures of the normal representation (hGPCR-lig: Table 2) and it is found in 74% of GPCR ligands in hGPCR-lig compared to only 26% of background compounds. Figure 2 shows some examples of GPCR ligands with this substructure overlaid, in sometimes unanticipated ways. These types of overlay also illustrate the completeness of coverage compared to the chemical fragment approach discussed in the Introduction. The other significant substructures in Table 2 are essentially variations of the first; the only differences are in number and length of carbon chains/atoms attached to the nitrogen atom. By scanning through the lists (e.g., Tables 1 and 2 and Tables 1 and 2 in the Supporting Information), a recurring theme becomes apparent. The topmost significant substructures are alkyl chains, some in combination with nitrogen, aromatic bonds, or combinations of these. Note that these lists represent substructures that occur often in GPCR ligands and not in the background molecules, thus being the distinguishing features that set these ligands apart from other organic compounds. In the 'normal' representation, a recurring theme is the alternating single/double bond feature, most likely being the substitute for aromatic bonds. Furthermore, the top significant substructures in this representation are alkylamines, chains of single-bonded carbon atoms, or combinations of both. The amine-containing substructures differ in number and length of bonded alkyl substituents; similarly, the length of the carbon tail differs as well as the position of the nitrogen within the tail. The substructure profiles for hGPCR-lig and GLIDA ligands were nearly identical, which implies that the method is stable for either data set. For instance, there was only one difference in the top 20 most significant substructures, and the ordering was virtually the same (compare Table 2 with Table 1 in the

Table 1. Best Discriminating Substructure for Each of the Four Elaborate Representations of the GPCR Ligands in hGPCR-lig ('Active' Column; Upper Line) and in the GLIDA Collection ('Active' Column; Lower Line)[a]

Representation	Substructure	Active	Control	P-Value
Normal		12,162 (74%) 16,434 (76%)	4,081 (26%)	7.35e-1707 3.46e-2138
Ar-Bonds		11,926 (72%) 16,142 (75%)	4,065 (25%)	7.25e-1611 3.04e-2026
Ar-Atoms/Bonds		8,627 (52%) 11,166 (52%)	1,977 (12%) 1,517 (9%)	1.06e-1354 1.50e-1760
Planar		13,050 (79%) 17,739 (82%)	5,534 (35%)	7.45e-1479 2.28e-1959

[a] For each representation, the same substructure is found in both databases, except for the 'aromatic atoms and bonds' representation. Here, the substructures differ in the presence of an extra atom label. The column labeled 'Control' refers to the background DIVERSet compound library. See Materials and Methods for an explanation of the p-value calculation.

Table 2. List of the 20 Most Frequent Substructures Found in GPCR Ligands (hGPCR-lig) Compared to the DIVERSet Background Compounds ('Normal' Representation)[a]

Nr	Substructure	Occurrence		P-Value
		GPCR ligands	Background compounds	
1		12,162 (74%)	4,081 (26%)	7.35e-1707
2		8,397 (51%)	1,273 (8%)	4.70e-1696
3		12,028 (73%)	4,128 (26%)	2.72e-1626
4		11,432 (69%)	3,741 (23%)	2.61e-1551
5		9,298 (56%)	2,096 (13%)	2.62e-1535
6		14,900 (90%)	7,844 (49%)	6.02e-1522
7		7,847 (48%)	1,225 (8%)	1.39e-1520
8		8,669 (53%)	1,900 (12%)	1.30e-1411
9		7,443 (45%)	1,178 (7%)	3.46e-1405
10		10,355 (63%)	3,152 (20%)	1.00e-1400
11		7,724 (47%)	1,539 (10%)	2.31e-1282
12		5,707 (35%)	594 (4%)	4.60e-1218
13		8,210 (50%)	2,039 (13%)	1.12e-1179
14		10,433 (63%)	3,761 (24%)	1.75e-1166
15		5,403 (33%)	534 (3%)	5.96e-1163
16		5,579 (34%)	666 (4%)	7.51e-1120
17		5,532 (34%)	712 (4%)	3.68e-1071
18		6,386 (39%)	1,147 (7%)	4.91e-1067
19		9,829 (60%)	3,559 (22%)	6.48e-1045
20		4,901 (30%)	516 (3%)	4.27e-1011

[a] Substructures were sorted according to significance, with the most significant substructure at the top. Thus substructure 1 (in bold) is found in 12,162 GPCR ligands (74%) compared to 4081 background compounds (26%). This finding is highly significant as judged from the corresponding p-value (7.35e-1707).

Supporting Information). This means that the most significant substructures cover the same area of chemical space in both the clinically promising candidates (hGPCR-lig) and compounds originating from scientific and patent literature (GLIDA). The topmost significant substructures are dominated by a few substructural themes that are common to the group as a whole. The hypothetical 'parent' fragment from which all frequent substructures derive would be an amine connected to an aromatic system through a carbon chain. This complies with the most common substructure, as defined by Sheridan et al.,[18] which was found in 21% of the ligands in the GPCR set. Again, the abundance of this substructure is probably due to the high number of aminergic receptor ligands present in the database (see also below). Even though the top significant substructures provide chemical insights found in the largest number of compounds, they also might reflect an obvious bias. A very simple one might be the commercial availability of reagents. Also, the hGPCR-lig database is largely filled with drug candidates that have reached the market or later-stage clinical trials. Due to the high attrition rates in drug discovery and development[45] these advanced compounds must have additional features for druglikeness that made them not fail beforehand, reducing the 'randomness' of substructure occurrence.

Hierarchical Partitioning To Distinguish GPCR Ligands from Other Ligands. For the first hierarchical analysis (the hierarchical partitioning), we used the hGPCR-lig and DIVERSet compounds. To find a new set of significant features, the experiment was repeated on subsets of the original sets: a subset in which all structures contained the substructure and another set in which they did not. The most significant substructure from the best performing representation was used to split the sets, and substructure mining was then repeated for the two subsets. The results are represented hierarchically as a tree (Figure 3). At the top of this hierarchy, we find the butyl amine substructure ('normal' representation) which occurs in 76% of GPCR ligands compared to 24% of background compounds. More than half of the molecules containing this aminergic tail also contain a substructure consisting of a heteroatom substituted with a methyl group and a propyl group. This substructure may overlay the butyl amine group or may be located elsewhere in the molecule. Molecules without the butyl amine substructure have a six-atom aromatic containing as most important structural feature. This aromatic chain is not closed to form a six-membered ring, meaning that this substructure is found not only in six-membered rings but also in aromatic systems containing fused five-membered rings etc. Since the 'normal' representation does not discriminate between aliphatic and aromatic bonds, it may lead to substructure contributions that are part of an aromatic ring system (Figure 2, e.g., compound IV). We therefore tested whether the use of a special aromatic bond type was the more suitable representation. Substructure occurrences for the aromatic bonds representation are listed in Table 3. When comparing the substructures of this representation with those of the 'normal' representation, a large overlap of the common motif was observed. For both representations, the most important moiety is the nitrogen substituted with one or more alkane chains. From these data, a second hierarchy (Figure 4) was constructed that did not consider the 'normal' representation. This approach resulted in a different substructure set than with the 'normal' representation. Compared to the amine from the 'normal' representation, the best-discriminating substructure now had a shorter tail attached to the nitrogen. This carbon tail is probably shorter due to the absence of contributing aromatic bonds that were represented as single bonds in the 'normal' representation. The first substructure with an aromatic bond was found at position 24 in the substructure list in Table 3, whereas in the 'aromatic atoms and bonds' representation the first aromatic substructure was

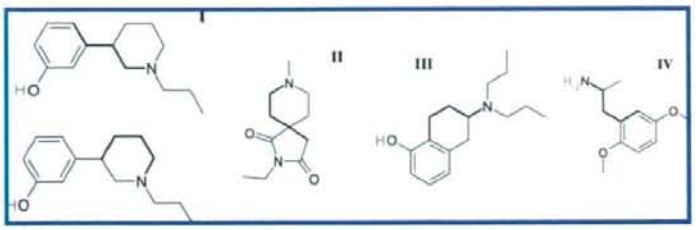


Figure 2. Example overlays on GPCR ligands of the most discriminative substructure, 'normal' representation. For the first ligand (I), two possible overlays are shown; multiple overlays are also possible for the ligands II and III. Note that the 'normal' representation uses Kekulé structures for aromatic systems and not separate types for delocalized bonds and aromatic atoms. This results in some interesting examples where the single bond of an aromatic ring is part of the aliphatic chain of the overlaid substructure, i.e. in structures I and IV.

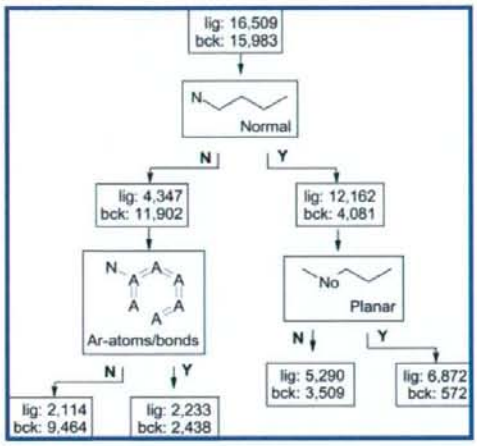
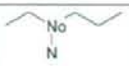
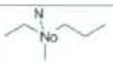
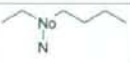
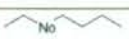
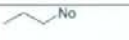
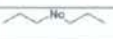
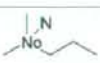
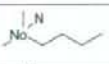
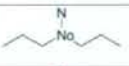
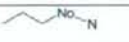


Figure 3. Hierarchical partitioning of GPCR ligands (hGPCR-lig-lig) versus background compounds (DIVERSet - bck). Ligands having the first substructure are taken out of the two sets (left path, N for 'No') or rather kept (right path, Y for 'Yes'), and the process of substructure mining is repeated. Numbers given are the absolute numbers of molecules used. Most discriminative substructures in any elaborate representation are shown. The name of the representation that yielded this substructure is given at the bottom of the box surrounding the substructures.

found much higher in the substructure list, i.e. at the third position. Iterating over the substructures with the 'aromatic atoms and bonds' representation leaves the impression that an aromatic ring is forming (Supporting Information: substructures 3 to 5 and 10 to 15 in Table 2). However, the fully closed ring is found much further down the list. The reason for this is that aromatic rings that differ in size contribute the same open substructures to the frequency tables. Since closed five-membered rings and closed six-membered rings cannot be overlaid, the support for these substructures is much lower. Note that even though the 'normal' representation may seem naïve, it still yielded the most significant substructure for GPCR ligands. Using this representation, a significant enrichment of the source sets can be accomplished, provided that bonds may be aliphatic or aromatic (Figure 1). Complementary to the analysis of typical substructures for the GPCR ligands was our analysis of substructures occurring more frequently in the DIVERSet background library. Thus, repeating the experiment for the background compounds yielded structural features that have low abundance in GPCR ligands. Table 4 gives an overview of the most-significant substructures for each representation. Similar as for the GPCR ligand analysis, a hierarchy was constructed for the background compounds (Figure 5). The tree is the same for analysis with and without the results of the 'normal' representation, since this representation did not produce any of the most significant substructures. The substructures that occurred often in the background set and not in the ligands are those that should be avoided when searching for GPCR ligands (since they seem to be related to inactivity). A carboxamide substructure was the most significant substructure found for the background (DIVERSet[37]) set. For this set, the aromatic bonds representation yielded the most significant finding (Table 4). Almost two-third of the compounds (depending on definition) in this set have a carboxamide or ester at the core of their scaffold, either linking two ring systems or linking a ring system with an aliphatic group. The high number of carboxamide and ester groups at the core of the molecules may reflect the simple organic reactions between alcohols and acids that have been used to construct the library. GPCR ligands differ not only in lower number of occurrence but also in position of these motifs. Where carboxamide and ester groups mainly form the linking groups between fragments in DIVERSet compounds, these motifs are also found as side-groups in GPCR ligands. This possibly reflects efforts to make drugs that are more soluble (for increased bioavailability) or to create prodrugs. The substructures in the GPCR hierarchy contained zero (Figure 4) or one heteroatom (Figure 3 and Figure 4), whereas those for the background set contained one, two, or three heteroatoms (Figure 5). This might reflect the lower number of hydrogen bond donors/acceptors in GPCR actives compared to inactives which was already noticed by Balakin et al.[31] Normally, as Feher and Schmidt pointed out,[46] the number of (small) heteroatoms is roughly equal in drugs as well as compounds in combinatorial libraries.

GPCR Subfamilies. We continued our analysis by focusing on individual classes of GPCR ligands. For the subfamily analysis, we continued with GLIDA and DIVERSet. For this second analysis, we derived smaller, sampled sets next to the two original full sets. The sampled sets, which were more convenient to work with, had the same substructure profiles as the full sets. For instance, the substructure lists of the full (Table 1 in the Supporting Information) and sampled set (Table 3 in the Supporting Information) of GLIDA are nearly identical. Substructures have the same order, e.g. only two shifts in positions occur in the 20 topmost structures for the 'normal' representation. Ligands were grouped hierarchically, based on the classification of the target GPCRs. The hierarchical grouping and levels where substructure analysis was performed are schematically presented in Figure 6. Similar to the analysis of GPCR ligands against

Table 3. Frequent Substructures for the GPCR Ligands (hGPCR-lig) Found with the 'Aromatic Bonds' Representation[a]

Nr	Substructure	Occurrence		P-Value
		GPCR ligands	Background compounds	
1		11,926 (75%)	4,065 (25%)	7.25e-1611
2		10,344 (65%)	2,828 (17%)	2.11e-1552
3		9,474 (59%)	2,257 (14%)	4.59e-1514
4		14,928 (93%)	8,258 (50%)	6.72e-1383
5		7,290 (46%)	1,241 (8%)	5.53e-1311
6		7,040 (44%)	1,168 (7%)	8.48e-1271
7		9,056 (57%)	2,444 (15%)	8.58e-1268
8		7,587 (47%)	1,574 (10%)	4.22e-1216
9		5,498 (34%)	525 (3%)	1.45e-1202
10		12,898 (81%)	6,091 (37%)	3.78e-1202
11		9,962 (62%)	3,286 (20%)	3.25e-1200
12		5,626 (35%)	606 (4%)	4.33e-1181
13		7,222 (45%)	1,467 (9%)	1.76e-1154
14		7,863 (49%)	1,868 (11%)	2.25e-1151
15		5,774 (36%)	736 (4%)	1.11e-1133
16		5,441 (34%)	590 (4%)	2.31e-1131
17		5,159 (32%)	514 (3%)	2.26e-1097
18		11,202 (70%)	4,916 (30%)	4.81e-997
19		8,939 (56%)	2,999 (18%)	2.10e-985
20		5,998 (38%)	1,110 (7%)	1.64e-966
...	...	...	...	...
24		10,819 (66%)	4,728 (30%)	2.04e-937

[a] The first occurrence of an aromatic bond is found in the 24th substructure. See for further explanation the legends of Tables 1 and 2.

a background set we now compared the ligands of a subgroup against all other ligands in the entire group. A group within a group is called a *subgroup*; the group that contains the subgroup is denoted as the *supergroup*. Substructures that have a significant preference for either the subgroup or the

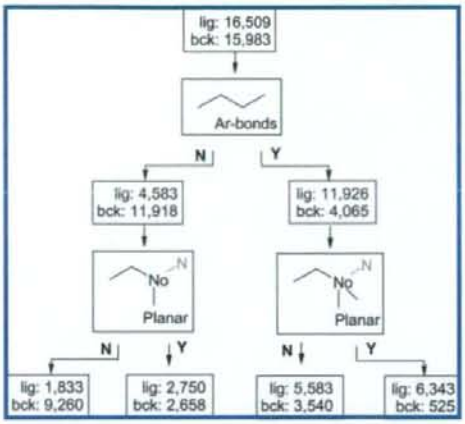
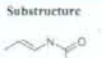
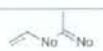


Figure 4. Hierarchical representation of the splits for GPCR ligands (hGPCR-lig) compared to the DIVERSet background collection, constructed without using the normal representation as in Figure 3. See the legend to Figure 3 for further explanation. As opposed to Figure 3, we now see a butyl chain without the amine group at the top of the hierarchy. Substructures containing the amine group are found one level down in the hierarchy.

Table 4. Best Discriminating Substructure Per Elaborate Representation for the DIVERSet Background Compounds Compared to the hGPCR-lig Collection[a]

Representation	Substructure	Background compounds	GPCR ligands	P-Value
Normal		6,247 (39%)	1,511 (9%)	9.03e-920
Ar-bonds		6,183 (39%)	1,434 (9%)	2.65e-938
Ar-Atoms & Bonds		6,998 (44%)	2,140 (13%)	1.92e-863
Planar		6,218 (39%)	2,101 (13%)	4.39e-658

[a] Note that the substructures do not have a geometric arrangement; the layout of double bonds and aromatic bonds is arbitrary.

supergroup are denoted as either *specific* or *avoiding*, respectively. Substructures that are specific for the supergroup are denoted as *generic*. Specific substructures are those that set ligands from one subgroup apart from ligands of the neighboring subgroups. Avoiding substructures are those that seem to avoid ligands of the subgroup but do occur in neighboring subgroups. Generic substructures are those that are common to a subgroup and the neighboring subgroups. Substructure lists not provided in this article are available as Supporting Information.

First, the differences between aminergic receptor ligands and all other GPCR ligands in GLIDA were analyzed (Figure 7). The 'planar ring' representation yielded the most significant substructures, even though substructures with parts of a planar ring were found at lower positions in the list. The best-discriminating substructures were methyl- and ethyl-substituted amines; the amine group of the endogenous ligands (e.g., dopamine, (nor)epinephrine, acetylcholine, etc.)